Submitting Electronic Data to CBER: Experience with SAS and JMP datasets

Ghanshyam Gupta, Ph.D.

Math Statistician, DBE, CBER

Peter A. Lachenbruch, Ph.D.

Branch Chief, Biostatistics, DBE, CBER

January 13, 1998

C B
E R

Types of Clinical Data

- **Vaccines data**
  - **Efficacy data (Thousands of subjects)**
  - **Titer data (50-100 subjects)**
- **Blood and Blood Products efficacy and safety data**
  - **twenty to hundreds per study**
- **Biological Therapeutics**
  - **Twenty patients to several thousand patients**

C B
E R

Mode of Data Submission -historical

- **SAS**
  - ***More than 90% of submissions***
- **ASCII files**
- **Spreadsheets: Lotus, Excel**
- **DBMS: Oracle, Access**
- **JMP**

Dataset Size

- 20 patients to 41000 patients
- 1 floppy disk to couple of CDs
  - under 1 Mb to 500 Mb
- No of studies: 1-50
- No of files: 1 - 300
- File size: under 1 Mb to 300 Mb

Example 1

- **GUSTO - Thrombolytic trial**
  - **One SAS file**
  - **About 80 Mb**
  - **More than 300 variables**
  - **41000 patients**

C B
E R

Example 2

- **Hepatitis A vaccine**
  - **200 subjects for a supplementary indication**
  - **<500 Kb**
  - **Disk submission**
  - **7 efficacy and safety variables**

Example 3

- **Hemophilia product**
  - **1 efficacy variable, multiple events per patient**
  - **32 patients**
  - **< 200 KB**

General Experience

- **Non-JMP and Non-SAS files**
  - **Dealt with on case by case basis - generally discouraged**
  - **Require extra time for manipulation**
  - **May need to be transported to other data analysis tools**
- **Always discuss with CBER before submitting such files.**

JMP Experience

- Some statisticians in CBER have used JMP for data analysis
- It has been used to review submissions by a few medical reviewers in CBER, experience is growing
- Viewing capabilities are a plus - SAS has none
- Data manipulation is less flexible than SAS, but certainly better than line listings

JMP Experience (Cont)

- Small files are satisfactory for data analysis, but it is not straightforward to manipulate multiple files
- JMP may not be able to open a large file on a PC because of limited memory. We have been unable to open some files even with 32MB machines
- Multiple large JMP files are difficult to work with.
- Multiple zipped JMP files may not open on a PC, again due to lack of resident memory on PC.

SAS Experience: Hardware

- Floppy disks - for small data sets
- Recently, larger SAS files have come on CD-ROMs
- E-IND uses only CD-ROM submission
- BLA/PLAs may submit on either 3.5” floppies or CD-ROMs - most have been CD-ROM
- Submit data files on separate CD-ROM (or floppy) from BLA/PLA narrative files.

SAS Experience: File Type

- We received files with the following extensions:
  - SAS files with .SD2 extension
  - SAS format library files with .SC2 extension
  - SAS program code files, generally with .sas extension
  - Other formats, etc. should be discussed with CBER

SAS Experience (Cont)

- **Zipped SAS files**
  - **If submission is on floppy disks, we have not had problems unzipping the files (assuming there is a limited number of files)**
  - **For many files on CDs, it is better not to zip them. Otherwise, we need to unzip all of them and save them on the PC which occupies too much space. If they are submitted unzipped, they can be used directly from CDs**
- **We prefer not to receive zipped files**

SAS Experience (Cont)

- **Zipped SAS transport files**
  - **Avoid submitting transport files unless needed due to completely different platforms (requires proc CPORT for conversion).**
  - **Additional steps required: (1) Unzip the transport files (2) Convert the transport to SAS files.**
  - **JMP Version 3.2 can read SAS files.**

Problems Experienced

- SAS files come with sponsor structure
  - c:\abc\def\ghi\klg\dgs\study1\
  - this makes it hard to find specific files
- Simplified structure
  - c:\studyname\datafiles
  - Format library file in the same directory
- Refer to the guidance document for preferred file naming conventions. File names should be 8 characters or less.

Problems Experienced (Cont)

- SAS data file names are not self explanatory. Provide details.
- Provide a file that discusses your naming conventions
- Example : File Name Definition

Data Definition Table

Table of File Name Definitions for the “Contrail” trial

Study Name/Number	File Name	File Description
Contrail 9701	AD_HX	Summary of medical history one record per patient
Contrail 9701	AE	Adverse Events multiple records per patient
Contrail 9701	Effic 1	Primary Efficacy Data one record per patient
Contrail 9702	Effic 2	Secondary Efficacy Data one record per patient

Problems Experienced (Cont)

- **Proc Contents for data dictionary is not sufficient. It simply provides variable type and label. It does not provide levels of categorical variables.**

Problems Experienced (Cont)

- For each variable within each dataset, provide sas-variable name, detailed description, type, format if any, values it can assume if categorical, and comments if any.
- Provide a file that details the variable names and definitions - This may be in tabular form as provided in the next example of data definition.

Dataset definition table

Variable	Description	Type	Format	Values
pt_ID	Five character pt number	C		
TxGRP	Assigned txmt grp	N	TX	1=p+200 mg 2=400mg 3=600mg
DOB	pt DOB	D		
Age	patient age in years	N		

Summary

- Minimize the number of steps needed to get to the data on a PC (we are migrating to PC only platform for the review process in CBER).
- If the number of files is small, do not zip them.
- Keep the format library file along with the data file(s)
- Provide a copy of the format file

Summary (Cont)

- If there are many files, provide them on CD-ROMs and do not zip them.
- Avoid transport files unless needed.
- Test your data submission. Be sure you can use the data on a standard desktoop PC.

Summary (Cont)

- Provide a detailed description of each SAS file.
- Provide a detailed sas-variable description.
- This information should be in a separate file in the submission.
- PROC CONTENTS is not sufficient.
- We anticipate that if datasets are used by the reviewers, Case Report Tabulations (i.e., line listings) will not be needed.